图书情报工作
第66卷 第2期 2022年1月

一种基于知识元变异的ESI研究前沿知识演进分析方法*

■ 孙震[1] 冷伏海[2]
[1] 山东理工大学信息管理研究院 淄博 255000 [2] 中国科学院科技战略咨询研究院 北京 100190

摘 要：[目的/意义]作为一类面向学科领域科技情报需求、针对全文本关键语义计量分析、旨在实现情报自动化到知识自动化实践应用的探索研究，本文基于语义标注和机器学习等技术，在前期研究从知识元共现视角探测研究前沿演进机理基础上，进一步提出一种基于知识元变异的研究前沿知识演进分析方法。[方法/过程]利用Word2vec词嵌入模型将知识元表示为词向量，通过计算知识元向量的欧几里得距离，利用K-means聚类方法识别具有相似语义语用关联的知识元簇集，计算历时簇集内各知识元TF-IDF值，对变异后知识元重要程度的突发变化结果进行定量测度，进而挖掘ESI研究前沿演进中的知识元变异特征和规律。[结果/结论]通过探测结果的对比检验发现，基于知识元变异的科学计量方法，不仅是对前期研究方法的补充和拓展，使得针对研究前沿内部知识运动规律的挖掘更加具体详实，更是在时间序列范畴内，能够尽早、及时探测研究前沿未来发展动向和关键情报信号的有力证据。
关键词：知识元 研究前沿 机器学习 全文本语义分析 钙钛矿太阳能电池
分类号：G250
DOI：10.13266/j.issn.0252-3116.2022.02.015

1 引言

当前，科学数据资源海量积累和增长，其中，科技文献作为记述科学发现和技术创新系统过程的知识载体，构成了比原始科学数据更加高效优质的文本大数据，更有助于科学知识发现及创新传播，但文本大数据的非结构化、弱语义表示和知识计算难度制约了科技文献的深层次、高效率利用能力[1]。因此，如何语义表示科技文献关键知识内容，开发细粒度知识单元关联计算方法，是当前科技情报面向学科领域智慧服务的重要工作[2-3]。

关键词和主题词虽能一定程度反映文章的研究主题及知识特征，但对于STEM（Science, Technology, Engineering & Mathematics）等领域科技文献来说，领域知识往往以关联知识单元形式密封在文献内，尤其是文献的Method/Experimental Section等“研究方法”部分[4]。如果能够限定某时段科技领域知识分布形态的文献数据范围，针对此类文献的特定知识单元，设计能够挖掘关键语义的知识单元关联计算方法，继而借助知识图谱等可视化方法，就可以展现该领域隐性知识分布形态，发现领域知识流动规律，实现潜在科学发现。具有对一定时段内世界科技前沿知识分布形态表象功能的数据集代表即ESI研究前沿（Research Fronts）数据库[5]。自2001年起，美国科学信息研究所ISI推出基本科学指标数据库ESI（Essential Science Indicators），并利用同被引分析方法进行研究前沿分析。“Research Fronts”，作为一个被定义为研究前沿的专业领域方法，即源自于科学研究间的某种共性，这种共性可能来自于实验数据，也可能源自科学假设、研究方法或科学概念，并反映在论文内科学家对其他科学家工作的引用这一学术行为。研究前沿记载了分散的研究领域的发生、汇集、发展等连续过程，在演进过程中，通过对研究前沿的施引文献分析，可以发现该领域的最新发展方向。

《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》中强调：“面向世界科技前沿、……，加强基础研究、注重原始创新，优化学科布局和研发布局。”在此背景下，围绕针对基础研究的ESI研究前沿数据，构建学科领域知识结构特征的语义加工方法，开展面向前沿的前瞻性战略

* 本文系国家社会科学基金项目“追踪研究前沿创新要素的领域知识元方法研究”（项目编号：21CTQ025）研究成果之一。
作者简介：孙震，馆员，博士；冷伏海，战略情报研究所所长，研究员，博士，博士生导师，通信作者，E-mail：lengfuhai@casipm.ac.cn。
收稿日期：2021-07-01 修回日期：2021-09-08 本文起止页码：136-148 本文责任编辑：杜杏叶

情报分析，既可以为科研人员把握研发动向、追踪前沿发展脉络、抢占制高点提供先机；也可以为国家及各级政府梳理学科战略布局侧重点、部署科技创新主攻方向、实现创新驱动发展提供决策支撑和智力支持。但当前针对 ESI“Research Fronts”的情报研究现状是，其常作为基础数据应用于“研究前沿”系列探测[6]、结合科学图谱探索国家表现[7]等工作，鲜有利用科技文献全文数据，从语义分析和知识计算视角探寻前沿内部知识结构变迁的情报实践。

基于上述背景，又由于本文前期研究已对学科领域知识元的概念内涵予以界定[8]、初步构建知识元计量方法的理论基础[9]，并已初步实证知识元方法之于 ESI 研究前沿知识演进的可行性和先进性[4]。因此，本文在前期研究基础上，进一步提出一种基于知识元变异的科学计量方法，将对 ESI 研究前沿的演进分析聚焦到群簇知识元层面，重点关注研究前沿演进过程中知识元的语义语用功能变化，事实上，这也是进一步从领域知识内容本身理解研究前沿演进机理的关键。前文[4]已经证明，借助情报实践的操作概念来理解，可将研究前沿看作具有语义语用功能的知识元集合，研究前沿演进的基础是与知识元相关的创新重组及应用变化，而对知识元共现链接关系的解读，虽是基于文本挖掘所抽取领域知识实体的分析，但本质上依赖的仍是对知识元计量结果的人工语义语用标注，只不过机器在此过程中完成了领域知识的自动集合和分类。本文在此背景下，仍以《2016 研究前沿》化学与材料科学领域“高效钙钛矿型太阳能电池”热点前沿为例，尝试利用机器学习技术自动识别知识元在研究前沿演进中的语义语用功能变化，探测相同语用环境下知识元群簇在研究前沿演进中所发生的突发变异现象，以期从不同视角、更加深入理解研究前沿演进时学科知识结构的变迁机理，也为面向学科领域的智慧型科技情报工作提供参考。

2 相关研究

科学文献知识特征可以划分为知识表现特征和知识实体特征。知识表现特征主要用于评价文献的学术影响力，知识实体特征则可分为外在知识实体特征和内在知识实体特征。外在知识实体主要是指文献表层的关键词、主题词等，常用于旨在促进知识发现的知识利用和转移研究，但基于外在知识实体特征的计量分析在科学文献的学科领域知识地图构建、捕捉学科领域思想、潜在内容关联发现等方面尚存在诸多局限[10]。近年来，已有许多学者关注基于内在知识实体特征的科学计量分析，并结合文献关联网络，在基于领域知识实体计量的科学知识发现方面取得了系列成果。

Y. Ding 等在 2013 年最早提出“实体计量学(Entitymetrics)”的概念，并从每篇文献中抽取基因、疾病、药物生物知识实体，构建实体引文网络，分析与 Metformin 相关研究的知识利用与转移方式[10]。此后，M. Song 等利用生物信息学 PubMed 种子文献及其参考文献，建立“基因 - 引文 - 基因”引用网络，检测基因间隐性的相互作用[11]。Q. Yu 等利用 PubMed Central 全文及其参考文献，抽取生物数据库知识实体，构建数据库链接网络，跟踪生物数据库的使用链接关系[12]。D. Lee 等以“阿尔茨海默病”为检索词，构建 4 种类型网络，从索引创建者、作者或引用者等不同观点视角捕获该领域的认知状态图景[13]。K. Lee 等通过层析内容分析法(tomographic content analysis)，将蛋白质、基因、MeSH 术语作为实体，探究文献内查询知识实体对异质知识实体网络的影响[14]。K. Li 等基于词典抽取每篇文献 R 程序包实体，构建单篇文献 R 程序包的共同提及网络(paper-level co-mention network)，探究 R 语言程序包在生物医学文献中的角色与使用状况[15]。

实体计量学将知识实体作为基本操作单位，真正实现了计量对象向文献内语义知识本体的深化，可更好的用于领域知识发现。遗憾的是，相较于生物医学领域，其他学科领域鲜有基于文献内领域知识实体计量分析的报道。究其原因，首先，从 PubMed 以外其他出版商很难获取全文数据，从全文数据获取到全文数据复用都面临困难；其次，生物医学领域知识实体关系构建往往限于全文 XML 文档和带有 PMID 的文章，机器读取处理较易，而其他科技领域全文常为付费下载的 PDF 格式，将其进一步转换为机器可读的 text 格式不但耗时，转换前后精确度也难以保证；再次，相较于生物医学领域词典的丰富完备，其他科技领域前沿交叉演化迅速、变迁方向多样，难以形成覆盖某领域全方向的知识实体词典，难以保证后期抽取处理的高准确率和召回率。

3 基础理论阐释

3.1 知识元变异理论

我国著名科学学家赵红州曾提出[16]：“任何一种科学创造过程，都是先把结晶的知识单元游离出来，然后再在全新的思维势场上重新结晶的过程。这种过程不是简单的重复，而是在重组中产生全新的知识系统，全新的知识单元。”相似地，将研究前沿看为一种复杂

的科学知识生态系统，研究前沿的演进看作科学知识思维重新结晶的过程，那么，在围绕前沿主题的特定知识范围内，伴随着前沿的演进变化，前沿内部也会发生知识元的离散和重组、演进和升华、衍生和转化，使研究前沿“知识结晶”形成一个从简单到复杂、从低级到高级的上升过程。在此期间，某些关键的知识元可能扮演着“知识基因”的角色，决定着特定领域知识的推进与突变。因为只有表征不同知识性质和形态的知识元，在不同前沿主题中迁移、引发知识元的重组、进而发生知识元结构的变异，才能改变前沿主题内部知识的链接和构造关系，以至于改变研究前沿的知识思维结晶状态、推动科学的创新和发展。

科学学奠基人贝尔纳（J. D. Bernal）认为[17]：“作为科学本身的要求，课题的形成和选择都是研究工作中最复杂的阶段。一般来说，提出课题比解决课题更困难，而评价和选择课题，便成为研究战略的起点。”反映到研究前沿中，研究前沿的创生往往集中于新产生的科学方向，新科学方向产生于新的科学选题，而新的科学选题来源于新的科学概念或认知。具体到钙钛矿太阳能电池等领域科技文献，能代表作者最初科学认知和文献核心科学概念的，便是科技文本中的“Methods”或“Experimental Section”等实验方法描述部分，因为如果科学家产生了一项最新的科学发现或科学发明，无论是技术革新升级还是新材料制备研发，均会在此部分进行详细阐述，以便于同行监督和科学实验重复。而如果某前沿主题在某时期形成，表现为科学现象就是一个新科学概念、科学发现、科学方法、科学技术、科学材料的出现，反映到文献中就是实验的材料组分、设备技术、操作方法的突变，映射到知识元的层面即是知识元发生的变异现象。意即，此时期的知识元生态系统，与上一时期知识生态系统相比，知识元构造的变异来源于表征某特定知识形态和特征的知识元的变化，前一时期时常出现的知识元成分被突然出现的知识元所替代。

3.2 ESI研究前沿中的知识元变异现象

从知识元变迁重组的视角来看，演变中的ESI研究前沿就包含了许多知识元变异现象。以钙钛矿太阳能电池前沿为例，透明导电玻璃基底、金属对电极、钙钛矿吸光层、电子传输层、空穴传输层是构成钙钛矿太阳能电池最重要的核心部件[18]，如果5种核心器件中有一种部件的材料组分发生变化，材料成分由一种换成了另一种（即表征研究材料的知识元发生变异），就会引发钙钛矿太阳能电池器件材料结构成分的重组，继而影响整体太阳能电池的光电转换效率和稳定性等特质。举例来说，在钙钛矿吸光层等其他器件材料组分均相同、制备温度也相同的情况下，如果仅将电子传输层材料由单一锐钛矿（anatase TiO_2）材料[19]、替换为锐钛矿和纳米纤维 TiO_2（anatase TiO_2 & nanofibers TiO_2）组成的复合物材料[20]，最终钙钛矿太阳能电池在短路电流密度、开路电压、填充因子和光电转换效率等关键性能方面表现均不相同，且大部分性能指标差异较大。这还仅为基于 TiO_2 同质材料的变异，如果将 TiO_2 替换为其他不同族类属别的化学材料，那么最终由表征研究材料的知识元组分重组而引发的钙钛矿太阳能电池性能改变将更大。

可见，ESI研究前沿知识元构造内部组分发生变异的信号，往往标志着由知识元重组而引发的科学知识重新结晶运动，并在一定条件下催发知识元所表征科学内涵的变化，进而推动科学要素重构，引起科学研究特质及性质的变革。

综上，本文将ESI研究前沿中钙钛矿太阳能电池领域的知识元变异具象化定义为：

定义1：将透明导电玻璃基底（Substrate）、金属对电极（Pole）、钙钛矿吸光层（Layer）、电子传输层（ETM）、空穴传输层（HTM）等构成钙钛矿太阳能整体器件的N种化学材料看为N元知识元组{MS, MP, ML, ME, MH, …, MN}，其在T1时刻文本的知识元构成为{MS1, MP1, ML1, ME1, MH1, …, MN1}，如果其在T2时刻文本至少有一种知识元材料成分发生变化，如由于电子传输层材料变化生成新的N元知识元组{MS1, MP1, ML1, ME2, MH1, …, MN1}，则说明发生了知识元变异现象。在此现象中，发生变异知识元所处的元组位置没有变化（对应到科学文献实验文本语料，其上下文位置知识元材料的组分构成和排列组合顺序没有变化），但就是因为处于同一位置、表征相同钙钛矿太阳能电池部件材料语义成分的变化，引发了电池整体性能和技术特征的变化。

4 研究方法

本文对知识元语义的描述主要基于计算语言学的分布假说（Distributional hypothesis）[21]。分布假说认为，词语的语义以及对词语语义的比较由其所处的上下文内容决定。本文所研究的知识元的“语义”是根据知识元的上下文（即上下知识元）以及所处的前沿主题范围予以判断，是指知识元进入交际后的意义（即知识元在不同语境下所代表化学材料的排布应用）。

事实上，对应前文[4]，当表征创新科学概念或创新材料的知识元演进变迁后，其对新前沿主题内其他知识元的影响实质是——使得其他知识元的上下文语境发生了变化（由同位置的知识元变化而引起组分排布变化）。因此，由于原始文本语料是抽取的每篇施引文献全文的实验部分，如果经过 POS 词性标注过滤器的清洗去噪，语料库中剩余文本的构造实质是带有上下文语境的“知识元词项袋”（Bag of Knowledge elements），该词袋中的知识元不是无序散乱分布，而是按科学家原始实验步骤的连续排列（continuous alignment）。伴随着研究前沿的演进，从钙钛矿太阳能电池关键器件材料到前期实验使用试剂溶剂的知识元构造成分会发生变异，而此时利用分布式语义（Distributional semantic）方法，恰好能够通过对知识元上下排列位置的神经网络学习，对知识元的语义语用加以表示和区分，如果在此基础上，再寻得一种可以对聚类后的知识元突发变异程度进行定量测度的指标，就可以清晰展现研究前沿内共时历时的知识元变异情况。

因此，本文首先利用基于分布假说构建的词语分散式表示工具 word2vec 词嵌入模型（Continuous Bag-Of-Words, CBOW 模型），基于上下文对知识元进行神经网络训练建模，将知识元表示为词向量，词向量是对具有相似上下文知识元的表示；然后，通过计算知识元向量的欧几里得距离（Euclidean distance）所构建的相似性矩阵，利用 K-means 聚类方法对其进行聚类，识别具有相似语义语用关联的知识元簇集，聚类即是对知识元变异运动结果的表征，聚类后的知识元簇具有相似的上下文，反映了具有相同语法语义特征的知识元组合，具有相同的语用功能，代表了钙钛矿太阳能电池前沿某种器件或材料的集合；最后，计算共时历时簇集内各知识元的 TF-IDF 值，对变异后知识元重要程度的突发变化结果进行定量测度，进而挖掘研究前沿演进中的知识元变异特征和规律。

具体研究方法流程如图 1 所示：

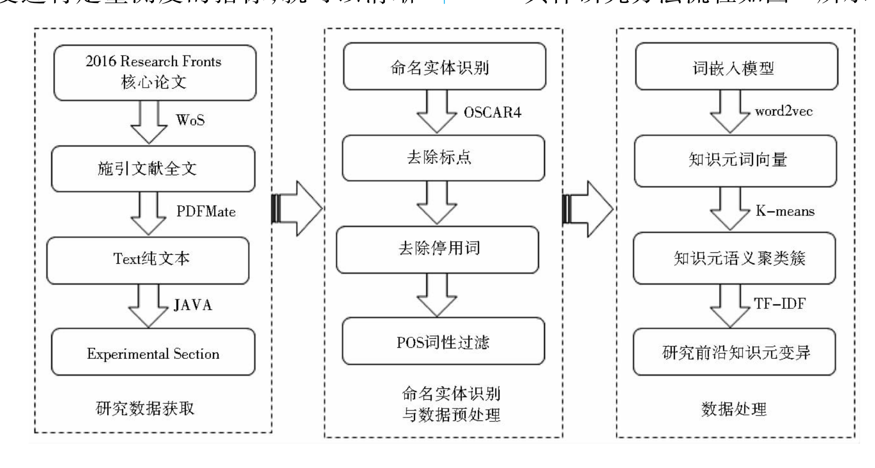


图 1　基于知识元变异的 ESI 研究前沿知识演进分析方法流程

4.1　Word2vec 词嵌入模型

本文对知识元上下构造随时间的变化研究，应用基于分布假说和神经网络的分布表示技术构建的 Word2vec 词嵌入模型进行探测。Word2vec 中包括两个模型：CBOW（Continuous Bag-Of-Words，连续词袋模型）和 Skip-Gram 两种模型，训练结果都是将语料中所有的词表示为相同维数分量为实数的连续向量。由于本文研究目标是求解预处理文本语料中上下文之间知识元的变化情况，而且截取的实验文本训练语料也并非大规模数据集合，因此使用 Word2vec 中的 CBOW 模型实施机器学习词向量训练。

4.2　K-means 聚类算法

利用 Word2vec 词嵌入模型计算的知识元词向量，是对实验文本中具有相似上下知识元位置成分的表征，即知识元的语用相似性（反映了知识元在钙钛矿太阳能电池中的材料器件应用）由知识元在实验文本中分布排列位置的相似性决定。而借助欧式距离对 Word2vec 知识元词向量的计算，则能够寻得具有相似上下文结构的知识元组分，且欧式距离越近、表明知识元间语义越相似，即知识元在研究前沿中的语用功能越相似，极有可能用于太阳能电池的同一种器件材料构成。

计算知识元词向量欧几里得距离所构建的相似性矩阵，利用 K-means 算法对其进行聚类，以识别具有相似语义语用特征的知识元簇集。聚类结果即是对知识元变异运动结果的表征，聚类后的目标知识元簇往往具有相似的上下文（在实验文本语料中具有相似的化学知识实体排列分布位置、具有相似的上下知识元词

项），簇集内的欧式距离往往较为相近，反映了具有相同语法语义特征的知识元组合，它们具有相同的语用功能、代表了钙钛矿太阳能电池前沿某类器件或材料的集合。相比于前文[4]中知识元关联及语义语用的人工标注，本文可以利用机器对相同语法、语义和语用功能的知识元实现自动识别并聚类。

4.3　TF-IDF 知识元突变度计量

知识元变异的特征实质，是某知识元所代表的化学材料成分，在某时期突然出现在某特定实验文本中，而与该知识元共同用来实验的其他化学试剂和材料并未改变（这些化学试剂和材料此时期在较多实验文本中均出现并广泛使用），也就是说，当发生知识元变异现象时，所对应文本和语词具有的规律特征是：发生变异的知识元词项在一篇实验文本中高频出现，且该知识元词项在同时期其他实验文本中出现的比例极小，意即该知识元词项此时期对于某特定前沿实验文本来说非常具有代表性和区分度，是特属于该特定实验文本的重要关键词项。而用以评估某词项对于一个文件集或语料库中一篇文本重要程度的统计方法，常用的正是 TF-IDF“词频 – 逆文件频率”算法，因此，本文用 TF-IDF 词项加权技术对各时期内同类簇知识元进行突发变异程度的表示和测算。

利用 TF-IDF 倾向于过滤掉某时期实验文本中应用广泛的常见知识元词项，保留此时期内突发变异程度较大、对特定实验文本较为重要的知识元词项，进而可以对具有同类化学材料属性知识元的共时历时突发变异程度进行定量测度。由于某时期某特定知识元可能在 n 篇实验文本中出现，会具有 n 个 TF-IDF 值，因此，为更好的表征该知识元在此时期的突发变异程度，本文将利用知识元在 n 篇文本中 n 个 TF-IDF 的平均值，作为该知识元的突发变异测度指标。即对于 t 时刻知识元词项 k，其突变度计算如公式 1 所示：

$$\mathrm{tf\ idf}_k = \frac{\mathrm{Sum}\ tfidf_{i,j}}{n} = \frac{\sum_{i=1}^{n} tfidf_{i,j}}{n} \quad （公式 1）$$

5　实证研究

5.1　Word2vec 知识元词向量训练

由于数据语料库规模大小会影响 Word2vec 知识元词向量的训练结果，机器学习的准确性也较为依赖神经网络输入层数据的量级，而且，为了与前文[4]结果进行更好的对比分析和延伸验证，本文沿用前文[4]对截取实验文本数据的时间标签分类，将对 ESI 研究前沿进行演化的语料库划分为 2010 – 2014 年（由于 2010 – 2013 年施引文献数据量过少，因此将其归并 2010 – 2014 年段分析）、2015 年、2016 年和 2017 年 4 个时间窗口。与前文[4]预处理方法相同，本文也将每个时期去除标点、去除停用词、N 元语词过滤预处理后的文本语料，在实施 OSCAR4 化学知识元实体识别后，利用 POS tagging（Part-of-Speech tagging）词性标注过滤器，过滤掉不含有 OSCAR4 化学实体标签的噪音数据，并最终经过 Notepad + + 等工具的进一步去重去噪，使得待处理的语料库仅为包含 OSCAR 化合物（chemical compound，CM）的知识元数据。但与前文[4]数据处理方法的不同在于，本文并未对预处理和去噪去重后的知识元文本进行 BOW 建模，而是直接导入 DeepLearning4J（DL4J）神经网络工具包，进行 Word2vec 词嵌入模型机器学习词向量训练。

应用 Word2vec 词嵌入技术，根据知识元的上下文信息，对知识元语义进行表示。每个时期原始语料为经过实体识别、预处理、POS 词性标注过滤之后，每篇文本语句被分割为一个二维列表，列表中的元素为文本处理后剩余的化学知识实体，这些化学实体知识元以字符串形式出现，表示如下：

Sentences = {[‘first’，‘knowledge element’]，[‘second’，‘knowledge element’]}，…

将这些带有原始实验上下位置排列分布顺序的知识元词项，导入 Word2vec 中的 CBOW 模型，就可以基于知识元的上下词项顺序，实施具有相同语义知识元的学习训练，预测所得知识元并非只有语义上的相似性，其体现的更是知识元间作为化学成分在实验中真实应用的关联，即与其他化学组分最终所生成制备材料的语用相关性。本文采用 Word2vec 模型的常用参数值，选取词向量维数为 100 维，输出的知识元词向量如表 1 所示：

表 1　知识元词向量示例

知识元	词向量（100 维）
TiO_2	（0.431 106 508，0.604 523 599，0.860 540 569，–0.063 856 05，…，0.225 055 814，0.109 347 574，–0.004 203 026，1.110 648 513）
Al_2O_3	（0.154 845 804，0.335 497 2，0.630 570 054，–0.097 430 952，…，0.060 130 555，0.137 183 696，0.055 556 033，0.409 101 218）
$CH_3NH_3PbBr_3$	（0.010 166 312，0.157 198 429，0.233 724 013，–0.004 123 966，…，0.027 948 08，0.082 523 182，0.074 739 113，0.052 901 864）
$CH_3NH_3PbCl_3$	（0.008 930 83，0.085 651 048，0.129 862 279，0.014 664 159，…，0.024 106 275，0.048 351 053，0.055 006 426，0.056 883 857）

注：选自 2010 – 2014 年段文本数据的词向量训练结果

5.2 K-means 知识元相似语义簇聚类

实施 K-means 算法最为首先、也是最为重要的步骤即是事先给定 k 值的选取,该初始聚类中心的选择对聚类结果有较大影响。依据肘部法则、轮廓系数等定量选取指标,结合前文[4]中知识元社区的数据分布结果,为保证聚类结果的准确性和最大收敛效应进行了 k 值选取的多次预处理实验,最终发现:针对不同时间窗口语料,当 k 值选取为 3 时,总体来看,不但处理得到的知识元语义簇聚类结果容易判读对比,且聚类收敛和分类效果均相对最佳。因此,对不同时间窗口下知识元词向量处理后的聚类语义簇均选定为 3 个,且 K-means 算法的最大迭代处理次数均设置为默认值 99 次。图 2 为对 2010 -2014 年知识元词向量进行 K-means 算法学习后的聚类结果分布图,为便于区分,将不同聚类簇的知识元用不同颜色和不同形状节点进行表示:聚类簇 1 中知识元节点为红色正方形;聚类簇 2 中知识元节点为绿色花型;聚类簇 3 中知识元节点为蓝色三角形。

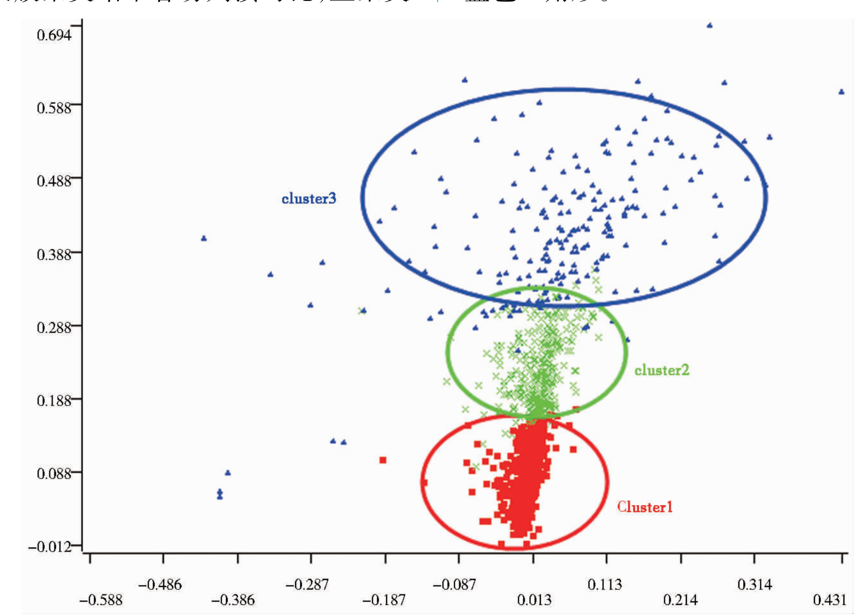


图 2 2010 -2014 年知识元词向量聚类分布结果

利用不同时段知识元 K-means 聚类的可视化结果,即可以通过各时期知识元语义语用相似性的收敛、发散、分布情况,直观判断研究前沿科学知识结构的稳定性和知识密集程度,进而对 ESI 研究前沿的内部知识流动规律有所把握。

5.3 TF-IDF 知识元突变度计量

Word2vec 通过知识元在实验中的上下位置排布,可以自动判别具有相似语义语用功能的知识元;而 K-means 的数据分割聚类功能,则可以实现对表征特定语义功能知识元对象的关联分类,使得每个组内部知识元的语义相关性较高,而组间知识元的语义相异性较高。在此之后,经过语义表征和语义分类处理后的知识元集合,就为某时期基于 TF-IDF 的知识元突发变异程度计量提供了天然数据集。

例如,图 2 中 2010 -2014 年段知识元语义聚类簇分布构成如表 2 所示:可以发现,Cluster1 和 Cluster2 中知识元的语用功能多表征钙钛矿太阳能电池前期实验制备常用的化学试剂或基础溶液,较少出现前文[4]中钙钛矿太阳能电池关键部件的化学材料组分;而 Cluster3 中知识元却多为构成关键器件材料的核心化学成分,如 $MASnX_3$、$CsSnI_3$、$MAPbI_3$ 等均可应用于钙钛矿太阳能电池吸光薄膜制备;而 In_2O_3、Sb_2S_3、NiO、ZnO、PbS 等则常应用于电子传输层、空穴传输层或支架阻挡层等核心器件。

表 2 知识元语义簇部分结果组成

知识元类簇	知识元组成
Cluster1	$BaTiO_3$; BFO; PDMS; DMF; Tb_4O_7; epoxy; $AgBiS_2$; FAI; Y-TiO_2; $AgSbS_2$; formaldehyde; C_3H_7NO; PANI; Zn_2TiO_4; PEI
Cluster2	Li-TFSI; IPFB; Sb_2Se_3; Ag_2S; hydrochloric; CQDs; acetic; CsI; Cu_2O; F8BT; phthalocyanine; FA; hydroiodic; dimethyl; CH_3NH_3Br
Cluster3	$CuSbS_2$; $MASnX_3$; MWCNT; DIO; In_2O_3; Sb_2S_3; $CsSnI_3$; MAP-bI_3; NiO; ZnO; PbS; ZrO_2; SnO_2; SiO_2; CdSe

注:选自 2010 -2014 年知识元词向量聚类簇,且簇内知识元按 TF-IDF 值降序排列

在此基础上,再借助 TF-IDF 突变度计算,就能够识别具有相似化学语义、表征相似化学材料语用功能知识元在某时期对于前沿主题文本的突发变异程度,

进而提早探得知识元对于该前沿领域未来创新发展的潜在影响效用。

例如，表 2 中 Cluster3 探得 2010 – 2014 年对于钙钛矿吸光薄膜材料突变度最大的知识元为 $MASnX_3$、$CsSnI_3$、$MAPbI_3$，在前文[4]知识元共现方法分析结果中，$MAPbI_3$ 在 2010 – 2014 年作为高共现率知识元被准确识别，但 $MASnX_3$、$CsSnI_3$ 由于共现频次较低，作为实验文本中低频词项无法予以识别，直至 2017 年才作为高共现知识元成对出现。而本文中，通过变异度计算，$MASnX_3$、$CsSnI_3$ 不仅在 2010 – 2014 年段就能被准确识别，且其在此时期变异程度较高，作为"知识地貌图"的突变"知识地势"，是未来可能影响技术创新方向的关键信号。如前文[4]所述，近年科学家致力于解决钙钛矿太阳能电池中有毒重金属 Pb 的环境污染问题，使得 $MASnX_3$、$CsSnI_3$ 等环境友好型无铅钙钛矿太阳能电池成为热点方向，这也为本文方法思路和上文判断提供了佐证。

可见，基于知识元变异的科学计量方法，不仅是对前文[4]知识元共现方法的补充和延伸，使得针对研究前沿内部知识运动规律的挖掘更加详实具体。更是在时序变迁下，能够尽早、及时探测研究前沿未来发展动向的有力情报证据。

5.4 基于知识元变异的 ESI 研究前沿演进分析

针对不同时间窗口下钙钛矿太阳能电池研究前沿截取的实验文本数据集，均经过 Word2vec 知识元词向量训练、K-means 知识元相似语义簇聚类、TF-IDF 知识元突变度计量等步骤，即可针对 ESI 研究前沿内部科学知识结构，由横向到纵向的描绘其随时间演进的知识流动及变化规律。为便于统一对照比较、且对每个时段研究前沿的知识变异特征进行更好的判读分析，在对不同时段文本数据进行处理时，每个时间窗口下最终均生成 3 个知识元语义聚类簇（为便于区分，不同聚类簇中的知识元节点用不同颜色和不同形状表示：聚类簇 1 中知识元节点为红色正方形；聚类簇 2 中知识元节点为绿色花型；聚类簇 3 中知识元节点为蓝色三角形），且每个簇集均展示 TF-IDF 变异度数值排名前 15 位知识元。

5.4.1 2010 – 2014 年研究前沿知识变异特征

ESI 研究前沿 2010 – 2014 年知识元语义簇聚类分布结果如图 3 所示，各簇内前 15 位高变异度知识元分布如表 3 所示：

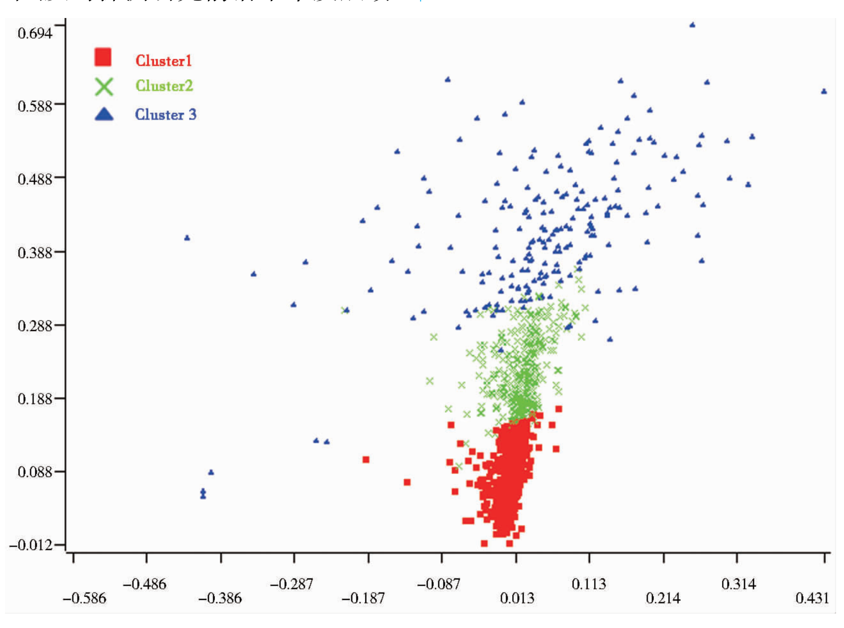


图 3　2010 – 2014 年研究前沿知识元语义聚类簇分布

可以看出，2010 – 2014 年作为钙钛矿太阳能电池研究萌发时期，此时知识元节点的聚类分布较为紧凑集中，各知识元语义聚类簇边缘分割也较为清晰。在语义语用功能方面，Cluster1 和 Cluster2 中拥有较高变异度的知识元，大都表征钙钛矿太阳能电池前期实验制备的化学试剂和基础溶液。Cluster3 则更多反映了此时期吸光层、电子传输层等钙钛矿太阳能电池核心器件材料的应用焦点。

Cluster1 聚焦制备阻挡层基底的聚二甲基硅氧烷（PDMS）、作为溶液溶解卤化物钙钛矿的二甲基甲酰胺（DMF）、制备支架材料的环氧树脂（Epoxy）、与其他材料合成制备纳米晶的银铋硫（$AgBiS_2$）[22]等。Cluster2

表 3　2010-2014 年研究前沿各语义簇知识元变异度分布

序号	Cluster1		Cluster2		Cluster3	
	知识元	变异度	知识元	变异度	知识元	变异度
1	$BaTiO_3$	0.529 9	Li-TFSI	0.479 4	$CuSbS_2$	0.316 3
2	BFO	0.517 0	IPFB	0.282 2	$MASnX_3$	0.285 8
3	PDMS	0.512 8	Sb_2Se_3	0.186 1	MWCNT	0.269 2
4	DMF	0.304 1	Ag_2S	0.135 5	DIO	0.215 1
5	Tb_4O_7	0.207 8	hydrochloric	0.132 5	In_2O_3	0.196 3
6	epoxy	0.193 3	CQDs	0.094 4	Sb_2S_3	0.128 8
7	$AgBiS_2$	0.187 0	acetic	0.080 3	$CsSnI_3$	0.087 3
8	FAI	0.158 2	CsI	0.073 6	$MAPbI_3$	0.084 1
9	Y-TiO_2	0.134 5	Cu_2O	0.068 0	NiO	0.065 2
10	$AgSbS_2$	0.133 6	F8BT	0.067 9	ZnO	0.057 3
11	formaldehyde	0.125 6	phthalocyanine	0.065 0	PbS	0.055 1
12	C_3H_7NO	0.123 8	FA	0.058 0	ZrO_2	0.054 1
13	PANI	0.123 8	hydroiodic	0.055 7	SnO_2	0.053 7
14	Zn_2TiO_4	0.123 8	dimethyl	0.055 1	SiO_2	0.051 3
15	PEI	0.117 2	CH_3NH_3Br	0.054 8	CdSe	0.047 9

聚焦常作为空穴传输能力提升剂的双三氟甲烷磺酰亚胺锂(Li-TFSI)、可抑制电池性能衰减的 IPFB、常用于溶解制备薄膜的盐酸(hydrochloric)、乙酸(acetic)等混合溶液。值得关注的是，Cluster1 中因钛酸钡($BaTiO_3$)、BFO($BiFeO_3$)均具优良铁电介电性质，$AgSbS_2$ 和 Zn_2TiO_4 共同具有敏化特性，Cluster2 中 Sb_2Se_3、Ag_2S、CQDs 常作量子点材料，这些知识元由于具有相同语用功能，不仅被准确识别处于同语义簇，且变异度指标数值均较高。Cluster3 中，铜锑硫($CuSbS_2$)太阳电池吸收层、$MASnX_3$ 和 $CsSnI_3$ 无铅钙钛矿吸光薄膜引起科学家关注，变异度指标均高于应用广泛的 $MAPbI_3$ 吸光材料；二碘辛烷(DIO)等也被发现对于器件性能改善有较大影响[23]。

5.4.2　2015 年研究前沿知识变异特征

ESI 研究前沿 2015 年知识元语义簇聚类分布结果如图 4 所示，各簇内前 15 位高变异度知识元分布如表 4 所示：

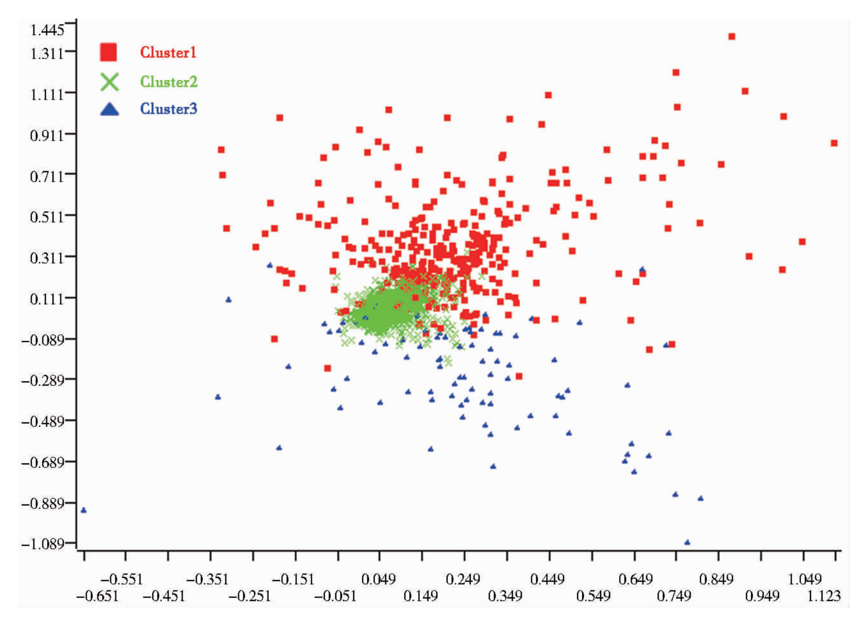


图 4　2015 年研究前沿知识元语义聚类簇分布

表 4 2015 年研究前沿各语义簇知识元变异度分布

序号	Cluster1		Cluster2		Cluster3	
	知识元	变异度	知识元	变异度	知识元	变异度
1	$Cs_3Sb_2I_9$	0.299 7	CQD	0.976 1	ZnO	0.076 3
2	Ag@ SiO_2	0.279 5	Zn_2SnO_4	0.289 3	FAI	0.066 9
3	pentane	0.259 4	CuS	0.280 6	P_3HT	0.059 8
4	BDT	0.245 0	$In(OH)_3$	0.278 1	PCBM	0.049 7
5	$ZnSnO_3$	0.173 6	PEO	0.267 3	$PC_{61}BM$	0.046 0
6	SiC	0.160 2	Bi_2S_3	0.252 3	tin	0.034 1
7	$CsSnI_3$	0.131 4	ODT	0.245 5	MABr	0.032 6
8	$MoSe_2$	0.129 5	C-PCBSD	0.205 3	Al_2O_3	0.031 9
9	$CsPbI_3$	0.123 9	MOF	0.168 8	bromide	0.030 7
10	MAPI	0.122 6	titanate	0.151 1	TiO_2	0.030 6
11	$FAPbI_3$	0.115 4	AgAl	0.138 9	MAI	0.030 5
12	PEDOT	0.102 6	PEI	0.138 2	GBL	0.029 6
13	Co_3O_4	0.102 3	GeO_2	0.133 0	IPA	0.028 0
14	PbS	0.099 8	mp-Al_2O_3	0.124 7	ITO	0.025 0
15	$MAPbI_2Cl$	0.093 6	$MASnI_3$	0.122 0	spiro-OMeTAD	0.023 4

2015 年研究前沿知识元聚类簇开始呈现交叉渗透迹象，聚类簇收敛中心的语义距离更为接近，各簇内知识元节点向四围扩散，说明此时进入研究前沿初步发展时期。科学家实验中使用化学材料种类更加丰富，因而知识元词向量映射的二维语义坐标距离间隔较近；语义聚类簇中心间隔走近，证明此时期科学家对太阳能电池具有相对较高效率和稳定性能的组件材料利用，初步形成共识。

语义语用功能方面，Cluster2 中知识元多为钙钛矿太阳能电池前期制备所需的新关注溶液试剂或中间产物修饰物。如，促进钙钛矿层结晶的锡酸锌(Zn_2SnO_4)[24]、常用作基于金属的碱性溶液 In(OH)3、可抑制反向电流的界面修饰材料聚氧化乙烯(PEO)[25]、用于衬底分子膜制备的疏水功能基团 ODT(正十八硫醇)、利于增强光吸收的空穴传输层掺杂物 MOF、用于修饰 ZnO 等电子传输层的富勒烯衍生物(C-PCBSD)等。

Cluster1 和 Cluster3 知识元节点较为分散，证明出现了许多新兴材料。其中，Cluster1 中知识元变异度数值明显高于 Cluster3，主要聚焦新型无机钙钛矿材料等方面：$Cs_3Sb_2I_9$、$CsSnI_3$、$CsPbI_3$ 等被证明具有合适带隙及高载流子迁移率；Ag 纳米相(Ag@ SiO_2 纳米颗粒)与 Al_2O_3 介孔层混合，被发现可明显提升钙钛矿复合薄膜光吸收性能[26]。Cluster3 则聚焦钙钛矿电子和空穴传输层材料：ZnO、Al_2O_3、TiO_2 成为代表性钙钛矿电子传输材料；FAI、MABr、MAI 引入前驱体制备高协调性钙钛矿层；需要说明的是，P_3HT 与 PCBM、$PC_{61}BM$ 一道用于空穴传输材料制备等引发科学家关注，3 种知识元变异度数值较为接近、且均高于应用广泛的 spiro-OMeTAD 传统材料，这种现象在本文中 2015 年即被准确发现，并识别出 $PC_{61}BM$ 这种潜在的相互作用化学成分，而前文[4]中直到 2016 年 P3HT 与 PCBM 才作为高频共现知识元对出现、成为研究热点。

5.4.3 2016 年研究前沿知识变异特征

ESI 研究前沿 2016 年知识元语义簇聚类分布结果如图 5 所示，各簇内前 15 位高变异度知识元分布见表 5。

2016 年研究前沿各语义聚类簇中知识元呈现更加分散的特点，各簇边界划分趋于清晰，簇间知识元覆盖现象减弱，各簇中心焦点距离变大，簇内知识元含有量也大幅增加，证明此时期进入钙钛矿太阳能电池前沿领域的快速发展时期。透过该现象也发现，借助对新型材料的研发设计，科学家此时期不只期望继续提高钙钛矿太阳能电池光电效率和稳定性，也旨在解决实验中面临的许多产业化问题，使其能尽早进入规模化生产应用。

在语义语用功能上，此时期又出现许多新的“知识地貌”突发信号。Cluster2 中知识元变异度数值均较高，主要聚焦用于增强制备性能的化合物及相关材料。其中，ALD-TiO_2、SAF-Ome、CuO-Cu_2O、In_2O_3-MWCNTs 等变异度数值相近知识元，多为具有增强制备性能作用的复合物：原子沉积(ALD)制备 TiO_2 可显著提高电池效率；SAF-OMe 空穴传输能力比 Spiro-OMeTAD 高三

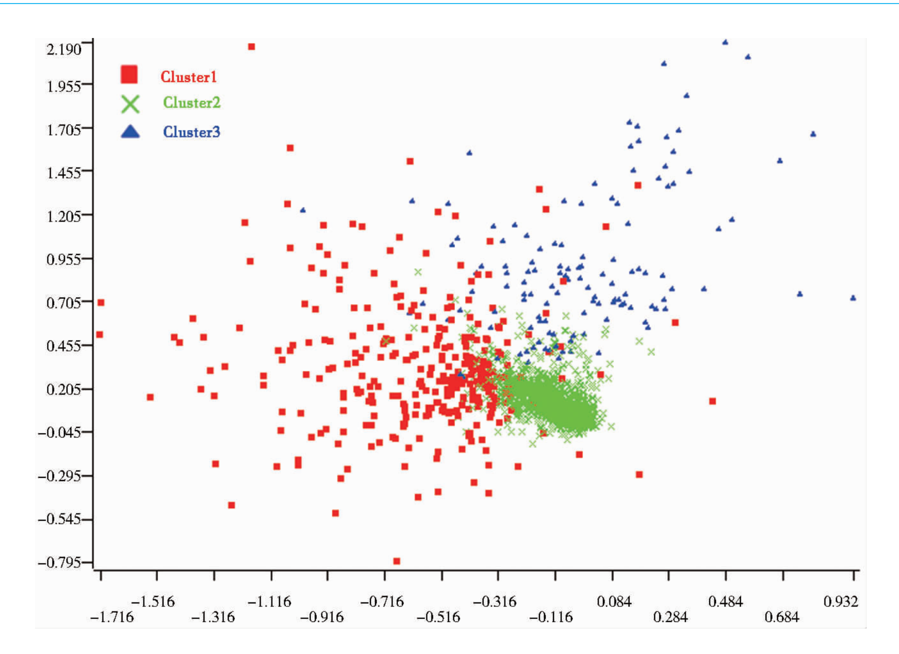


图 5　2016 年研究前沿知识元语义聚类簇分布

表 5　2016 年研究前沿各语义簇知识元变异度分布

序号	Cluster1		Cluster2		Cluster3	
	知识元	变异度	知识元	变异度	知识元	变异度
1	CQD	0.441 7	ALD-TiO_2	0.499 5	$HPbI_3$	0.323 6
2	MnO_2	0.322 0	SAF-Ome	0.499 5	HOIP	0.245 9
3	CZTSe	0.281 3	CuO-Cu_2O	0.444 6	GO	0.109 1
4	MoS_2	0.269 4	In_2O_3-MWCNTs	0.408 7	BiI_3	0.079 7
5	NiS	0.263 2	AgCuS	0.360 5	PEG	0.076 5
6	Bi_2Te_3	0.244 4	$AgBiS_2$	0.331 6	NiO	0.070
7	PCBMs	0.233 3	$BaSi_2$	0.328 4	PDMS	0.068 5
8	WSe_2	0.204 3	ZSO	0.312 6	SnI_2	0.067 35
9	CTAB	0.188 6	NH_4SCN	0.296 9	ZnO	0.059 4
10	$BaZrS_3$	0.171 7	$BiFeO_3$	0.295 1	$FAPbI_3$	0.059 0
11	nitrides	0.133 6	AgNWs	0.290 2	PVP	0.058 1
12	germanium	0.121 5	Cs_4PbBr_6	0.281 0	MABr	0.048 8
13	$MA_3Bi_2I_9$	0.113 1	NGQDs	0.266 7	Al_2O_3	0.041 5
14	$MASnI_3$	0.074 1	C60-SAM	0.257 8	SnO_2	0.039 0
15	CdSe	0.073 6	$CsBi_3I_{10}$	0.245 8	THF	0.036 2

倍以上[27]；CuO-Cu_2O 半导体纳米棒阵列可有效催化光电合成反应。而 AgCuS、$AgBiS_2$、$BaSi_2$、Zn_2SnO_4（ZSO）、$BiFeO_3$、Cs_4PbBr_6、$CsBi_3I_{10}$ 等系列光敏半导体和纳米晶材料，则常应用于太阳能薄膜制备。

Cluster1 和 Cluster3 聚类簇中知识元节点更加散布，节点距离间隔较远，知识元间语义语用类别差距相对较大。其中，Cluster1 多关注钙钛矿吸光层掺杂制备材料：如可作为低电阻金属氧化物掺杂制备光吸收层多孔骨架的 MnO_2、掺杂钙钛矿吸光层可有效提高光电及稳定性的 MoS_2、可用于制备超薄柔性太阳能电池的 WSe_2 等。Cluster3 多为实验常用基础化学试剂，但也不乏重要前沿信号：如可简化钙钛矿薄膜合成修复制备工艺的 $HPbI_3$[28]、可改善器件光伏性能的聚乙二醇（PEG）、可显著提高光电效率的聚乙烯吡络烷酮（PVP）[29]等，均成为钙钛矿太阳能电池研究此时的重要关注点。

5.4.4　2017 年研究前沿知识变异特征

ESI 研究前沿 2017 年知识元语义簇聚类分布结果如图 6 所示，各簇内前 15 位高变异度知识元分布如表 6 所示：

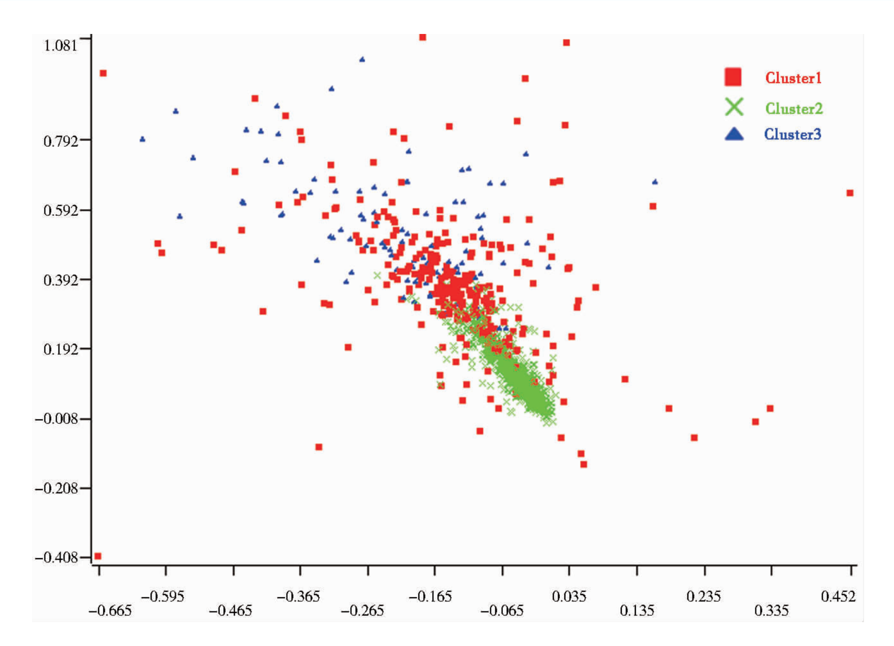


图 6 2017 年研究前沿知识元语义聚类簇分布

表 6 2017 年研究前沿各语义簇知识元变异度分布

序号	Cluster1		Cluster2		Cluster3	
	知识元	变异度	知识元	变异度	知识元	变异度
1	a-$FAPbI_3$	0.589 7	GeP	0.917 3	ZnO	0.069 4
2	SWCNT	0.493 1	CuS	0.469 9	$MASnI_3$	0.064 6
3	$KCoF_3$	0.279 2	MoS_2	0.319 2	$PC_{61}BM$	0.051 6
4	$RbCoF_3$	0.272 5	AgNCs	0.314 5	dimethyl	0.049 6
5	N-TiO_2	0.232 9	OLCNS:Ag	0.284 3	diethyl	0.049 1
6	$FASnI_3$	0.227 8	PbSe@ CdSe	0.259 6	isopropoxide	0.048 2
7	h-BN	0.223 0	CIGSSe	0.253 5	acid	0.043 4
8	$MCoF_3$	0.212 7	Y_2O_3	0.234 2	acetate	0.041 9
9	$Cs_2InAgCl_6$	0.176 6	Cu_2S	0.223 1	SnI_2	0.039 3
10	$AgBiI_4$	0.153 8	rubrene	0.205 2	SnO_2	0.036 0
11	PVK	0.137 6	TPBC	0.203 9	TiO_2	0.034 6
12	$Cs_2[AgIn]Br_6$	0.134 6	MWCNTs	0.196 6	SnF_2	0.032 9
13	$Cs_2InBiCl_6$	0.118 6	NWs	0.190 6	PCBM	0.031 2
14	RbCsMAFA	0.117 8	SWNTs	0.185 5	ethyl	0.031 1
15	$RbSnI_3$	0.112 9	SWCNTs	0.176 2	isopropyl	0.030 6

2017 年研究前沿各语义簇呈现更加交叉靠拢、越发覆盖渗透的态势,各簇内知识元分布出现明显收拢聚焦迹象,各簇聚类中心间隔也进一步缩小,说明此时期研究前沿进入稳定发展阶段。具有相似语义语用功能的知识元材料种类明显增多,构成钙钛矿太阳能电池的核心器件材料趋于稳定,各簇间知识元材料在实验中化学作用归类也不似往年那么边界明晰,但各簇内知识元材料在现实中的实验应用目的却更为相近。

Cluster1 和 Cluster3 知识元分布形态较为相似,只是 Cluster3 相对更趋于收敛,知识元间语义相似程度更高,其相似语用功能聚焦在半导体电子传输材料等方面:如同为金属氧化物半导体材料的 ZnO 和 TiO_2;同为富勒烯衍生物且常用于电子传输材料的 $PC_{61}BM$ 和 PCBM;制备 $MASnI_3$ 无铅清洁太阳能电池的 SnI_2、SnO_2、SnF_2 等。Cluster1 知识元变异度数值明显高于 Cluster3,具有许多推动器件性能革新的重要信号:西安交通大学吴朝新团队实现高效柔性非铅甲脒锡碘($FASnI_3$)钙钛矿太阳能电池,引发关注[30];通过梯度带隙提高太阳光谱利用是叠层电池的未解决难题,科学家此时提出一种利用六方氮化硼(h-BN)作为中间单层形成梯度带隙的方法,引起轰动[31]。

Cluster2 中知识元排布则较为紧凑,语义语用功能

也更为相似，主要聚焦在新型半导体晶体光敏材料和碳纳米系列材料等方面。例如，GeP、CuS、MoS_2、PbSe@CdSe、CIGSSe、Y_2O_3、Cu_2S 等半导体晶体光敏材料，常作为量子点、染料敏化、多结太阳能电池等相关器件材料；AgNCs（银纳米簇）、OLCNS:Ag（洋葱状碳纳米球复合银）、MWCNTs（多壁碳纳米管）、NWs（纳米线）、单壁碳纳米管（SWNTs/SWCNTs）等知识元，则体现了科学家此时期对于碳纳米系列材料的应用关注[32]。

6 结语

作为一类面向学科领域情报需求、基于全文本分析和关键语义计算的情报实践探索，本文首先基于分布假说理论，使用 Word2vec 模型训练知识元语义；然后，利用 K-means 聚类学习算法，寻找知识元群簇的语义语用分类和相互作用关系；最后，通过同语义簇知识元的 TF-IDF 数值，计算其相对前沿主题文本的突发变异程度，进而探测时序变迁下 ESI 研究前沿的知识演进特征和可能引发创新的关键信号。

通过邀请领域专家检验和专业学科文献查证等途径，发现本文方法能较好识别各时期可能推动前沿创新发展的关键情报信号，前文[4]分析结果也恰好成为对本文结果的检验。事实上，前文[4]基于知识元共现的识别结果更多的是反应达到一定热度的前沿热点方向，而本文结果实质为各时期“知识地貌”图中突现的“知识势场”信号，更可能为前沿演进中知识迁移的关键转折节点，是刚冒头的前沿方向。本文所识别信号往往不是钙钛矿太阳能电池大部件的整体革新，更为常见的是基于与太阳能电池核心器件相关的微小改良和升级（或者化学试剂的添加掺杂等微小实验步骤）而引发的整体性能提升。其实，由微小改进改良到整体器件性能质的提升过程，才真正体现了科学在现实中的发展轨迹——由点到面、由小的关键突破推动整体的科技创新研发。

还需要说明的是，本文所提理论方法与技术方案，虽仅以 ESI 研究前沿数据作为案例，但整套思想和方法设计并不局限于 ESI 研究前沿，其对于利用引文关系、语词关系和其他计量指标所识别的传统“研究前沿”仍具有较强普适性和可借鉴推广价值，因此，未来也将有针对性地开展本文方法与技术方案对于其他研究前沿数据及演进规律的挖掘分析。

参考文献：

[1] 孙坦. 图书馆智能知识服务的未来[J]. 中国图书馆学报, 2021, 47(2): 15 - 18.

[2] 宋宁远, 裴雷, 王春迎. 科学论文语义增强的研究进展与趋势研判[J]. 图书情报工作, 2021, 65(1): 82 - 90.

[3] 曾建勋. “十四五”期间我国科技情报事业的发展思考[J]. 情报理论与实践, 2021, 44(1): 1 - 7.

[4] 孙震, 冷伏海. 一种基于知识元共现的 ESI 研究前沿知识演进分析方法[J]. 情报学报, 2018, 37(11): 1095 - 1113.

[5] 冷伏海, 孙震, 周秋菊.《2015 研究前沿》报告的研制实践与相关探讨[J]. 智库理论与实践, 2016, 1(2):79 - 87.

[6] Clarivate Analytics. Clarivate and the Chinese Academy of Sciences release annual joint report to identify 100 + research fronts[EB/OL]. [2020 - 11 - 13]. https://clarivate.com/news/clarivate-and-the-chinese-academy-of-sciences-release-annual-joint-report-to-identify-100-research-fronts/.

[7] 王小梅, 邓启平, 李国鹏, 等. ESI 研究前沿的科学图谱及在纳米领域的应用[J]. 图书情报工作, 2017, 61(12): 106 - 112.

[8] 孙震, 冷伏海, 张晋辉. 基于知识元的科学计量方法及其实证研究[J]. 图书情报工作, 2017, 61(23): 89 - 99.

[9] 孙震, 冷伏海. 基于知识元的新型科学计量范式探析[J]. 情报学报, 2017, 36(6): 555 - 564.

[10] DING Y, SONG M, HAN J, et al. Entitymetrics: measuring the impact of entities[J]. Plos one, 2013, 8(8): e71416.

[11] SONG M, HAN N G, KIM Y H, et al. Discovering implicit entity relation with the gene-citation-gene network[J]. Plos one, 2013, 8(12): e84639.

[12] YU Q, DING Y, SONG M, et al. Tracing database usage: detecting main paths in database link networks[J]. Journal of informetrics, 2015, 9(1): 1 - 15.

[13] LEE D, KIM W C, CHARIDIMOU A, et al. A bird's-eye view of alzheimer's disease research: reflecting different perspectives of indexers, authors, or citers in mapping the field[J]. Journal of Alzheimer's disease, 2015, 45(4): 1207 - 1222.

[14] LEE K, KIM S Y, KIM E H J, et al. Comparative evaluation of bibliometric content networks by tomographic content analysis: an application to Parkinson's disease[J]. Journal of the Association for Information Science and Technology, 2017, 68(5): 1295 - 1307.

[15] LI K, YAN E. Co-mention network of R packages: scientific impact and clustering structure[J]. Journal of informetrics, 2018, 12(1): 87 - 100.

[16] 赵红州, 蒋国华. 知识单元与指数规律[J]. 科学学与科学技术管理, 1984 (9): 39 - 41.

[17] 贝尔纳. 科学研究的战略[C]// 科学学译文集. 北京: 科学出版社, 1980.

[18] 姚鑫, 丁艳丽, 张晓丹, 等. 钙钛矿太阳电池综述[J]. 物理学报, 2015, 64(3): 135 - 142.

[19] BURSCHKA J, PELLET N, MOON S J, et al. Sequential deposition as a route to high-performance perovskite-sensitized solar cells[J]. Nature, 2013, 499(7458): 316 - 319.

[20] DHARANI S, MULMUDI H K, YANTARA N, et al. High efficiency electrospun TiO_2 nanofiber based hybrid organic-inorganic perovskite solar cell[J]. Nanoscale, 2014, 6(3): 1675 - 1679.

[21] HARRIS Z S. Distributional structure[J]. Word, 1954, 10(2/

3): 146 - 162.
[22] CHEN C, QIU X, JI S, et al. The synthesis of monodispersed $AgBiS_2$ quantum dots with a giant dielectric constant[J]. CrystEngComm, 2013, 15(38): 7644 - 7648.
[23] LIANG P W, LIAO C Y, CHUEH C C, et al. Additive enhanced crystallization of solution-processed perovskite for highly efficient planar-heterojunction solar cells[J]. Advanced materials, 2014, 26(22): 3748 - 3754.
[24] BERA A, SHEIKH A D, HAQUE M A, et al. Fast crystallization and improved stability of perovskite solar cells with Zn_2SnO_4 electron transporting layer: interface matters[J]. ACS applied materials & interfaces, 2015, 7(51): 28404 - 28411.
[25] 董豪鹏. 有机无机杂化钙钛矿成膜前的界面修饰及其器件光伏特性[D]. 北京: 清华大学, 2015.
[26] 马东超. 银纳米相吸收增强型 $CH_3NH_3PbI_3$ 钙钛矿薄膜及电池的制备研究[D]. 哈尔滨:哈尔滨工业大学, 2015.
[27] WANG Y K, YUAN Z C, SHI G Z, et al. Dopant-free spiro-triphenylamine/fluorene as hole-transporting material for perovskite solar cells with enhanced efficiency and stability[J]. Advanced functional materials, 2016, 26(9): 1375 - 1381.
[28] PANG S, ZHOU Y, WANG Z, et al. Transformative evolution of organolead triiodide perovskite thin films from strong room-temperature solid-gas interaction between $HPbI_3$-CH_3NH_2 precursor pair[J]. Journal of the American Chemical Society, 2016, 138(3): 750 - 753.
[29] 李建丰, 赵创, 张恒, 等. 利用 PVP 添加剂提高钙钛矿太阳能电池光伏性能[J]. 发光学报, 2016 (1): 56 - 62.
[30] XI J, WU Z, JIAO B, et al. Multichannel interdiffusion driven $FASnI_3$ film formation using aqueous hybrid salt/polymer solutions toward flexible lead-free perovskite solar cells[J]. Advanced materials, 2017, 29(23): 1606964.
[31] ERGEN O, GILBERT S M, PHAM T, et al. Graded bandgap perovskite solar cells[J]. Nature materials, 2017, 16(5): 522.
[32] LI Z, DONG J, LIU C, et al. Improved optical field distribution and charge extraction through an interlayer of carbon nanospheres in polymer solar cells[J]. Chemistry of materials, 2017, 29(7): 2961 - 2968.

作者贡献说明:
孙震:设计论文框架思路,数据收集、整理与分析,撰写和修改论文;
冷伏海:提出研究命题和主体思想,审校和修改论文。

An ESI Research Fronts Knowledge Evolution Analysis Method Based on Knowledge Element Variation

Sun Zhen[1] Leng Fuhai[2]
[1] Institute of Information Management, Shandong University of Technology, Zibo 255000
[2] Institutes of Science and Development, Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] As an exploratory research, this paper is oriented to the needs of scientific and technological information in the specialized discipline domain, and aims to realize the quantitative analysis of key semantics of the full text and the practical application shift from "information automation" to "knowledge automation". On the basis of previous studies from the perspective of knowledge element co-occurrence to explore the evolution mechanism of ESI research fronts, this paper further proposes a research front knowledge evolution analysis method based on knowledge element variation. [**Method/process**] Firstly, knowledge elements were represented as word vectors by word2vec word embedding model. Then, this paper calculated Euclidean distance of knowledge element vectors, and identified knowledge element clusters with similar semantic and pragmatic association by K-means clustering method. Finally, TF-IDF values of each knowledge element in the diachronic cluster were calculated. Through the quantitatively measurement of sudden changes in the importance of knowledge elements, the characteristics and rules of knowledge element variation were mined in the process of ESI research fronts evolution. [**Result/conclusion**] Through the comparative test of the detection results, it is found that the scientometric method based on knowledge element variation is not only a supplement and expansion of the previous research methods, but also makes the mining of the internal knowledge movement law of ESI research fronts more specific and detailed. Moreover, in the scope of time series, it is a strong evidence that the future development trend and key information signals of the ESI research fronts can be detected as soon as possible.

Keywords: knowledge element research fronts machine learning full-text semantic analysis perovskite solar cell